BLPA : Bayesian Learn-Predict-Adjust Method for Online Detection of Recurrent Changepoints

Maslov, Alexandr

Institute of Electrical and Electronics Engineers
2017

Maslov , A , Pechenizkiy , M , Pei , Y , Zliobaite , I , Shklyaev , A , Kärkkäinen , T & Hollmén , J 2017 , BLPA : Bayesian Learn-Predict-Adjust Method for Online Detection of Recurrent Changepoints . in IJCNN 2017 : The International Joint Conference on Neural Networks . Proceedings of ... International Joint Conference on Neural Networks , Institute of Electrical and Electronics Engineers , Piscataway, NJ , pp. 1916-1923 , International Joint Conferenceon Neural Networks , Anchorage , United States , 14/05/2017 . https://doi.org/10.1109/IJCNN.2017.7966

http://hdl.handle.net/10138/307576
https://doi.org/10.1109/IJCNN.2017.7966085

unspecified
acceptedVersion

BLPA: Bayesian Learn-Predict-Adjust Method for Online Detection of Recurrent Changepoints

Alexandr Maslov*, Mykola Pechenizkiy*, Yulong Pei*,
Indrė Žliobaitė§, Alexander Shklyaev‡, Tommi Kärkkäinen† and Jaakko Hollmén§

* Eindhoven University of Technology, Dept. of Mathematics and Computer Science, The Netherlands
Email: {a.maslov, m.pechenizkiy, y.pei.1}@tue.nl

† University of Jyväskylä, Department of Mathematical Information Technology, Finland
Email: tommi.karkkainen@jyu.fi

‡ Lomonosov Moscow State University, Dept. of Mechanics and Mathematics, Russia
Email: ashklyaev@gmail.com

§ Aalto University, Dept. of Computer Science, Finland
Email: zliobaite@gmail.com, Jaakko.Hollmen@aalto.fi

***Abstract*—Online changepoint detection is an important task for machine learning in changing environments. Presence of noise that can be mistaken for real changes makes it difficult to develop an effective approach that would have a low false alarm rate and being able to detect all the changes with a minimal delay. In this paper we study how performance of popular Bayesian online detectors can be improved in case of recurrent changes. Modeling recurrence allows us to anticipate future changepoints and predict their time locations. We propose BLPA, an efficient approach for inducing and integrating recurrence information in the streaming settings, and demonstrate its effectiveness in the experimental study on synthetic and real-world datasets.**

I. Introduction

Online change detection is practically relevant in many domains, such as medicine, energy production, industrial processes monitoring [1]. In machine learning and data mining research areas change detection is often studied in the context of problem of concept drift happening due to changes in the underlying data distribution over time [2]. A popular approach for handling concept drift is to monitor data or model performance for changes and to adapt model using most recent data collected after the last detected change [3].

In this paper we consider a change detection task in a one-dimensional univariate time series data streams. Further in the text we denote a univariate vector of observations either as $\langle x_i \rangle_{i=1}^{n}$ or as $\boldsymbol{x}_{1:n}$, i.e.

$$\boldsymbol{x}_{1:n} \equiv \langle x_i \rangle_{i=1}^{n} \equiv \langle x_1, \ldots, x_n \rangle$$

Input to the change detector is a vector of observations $\langle x_t \rangle$ indexed by the timestamps $t \in \mathcal{T}$. Timestamps is an ordered vector of time moments $\mathcal{T} \equiv \langle t_1, \ldots, t_T \rangle$ when observations were taken with a constant sampling rate. *Changepoint is a time moment when statistical properties of the data stream change significantly according to the predefined criteria.* Changepoint is identified by the moment of time when it happened (further - 'time location of the change'). The sequence of changes is denoted as $\langle c_i \rangle_{i=1}^{k} \in \mathcal{T}$ and an individual change from this sequence as c_i. Changes should be detected online when the only information observed until current moment of time can be used for an analysis.

The top plot (A) in the Fig. 1 illustrates an example of the input signal with three changepoints in the mean value at the moments $\boldsymbol{c}_{1:3} = \langle 5, 10, 14 \rangle$. Change is usually detected with some time delay δ. The change detection task is to detect changes $\boldsymbol{c}_{1:3}$ with as small a possible delay δ while not alarming changes at any other time moments, i.e. avoiding false alarms as much as possible.

An event when the change was alarmed by the detector while there is actually no change is called False Positive (FP). Outliers and noisy changes in the input signal may cause FPs.

While the majority of existing change detection techniques focus on individual changepoint detection and assume that changepoints are not predictable, Fig. 1 illustrates use cases in which changes are expected to reappear over time. In this paper we focus on such setting, addressing the problem of detecting changes in noisy signals with recurrent changes.

Our approach (called **BLPA** method) is based on the hypothesis that if probability distribution of the time intervals between changepoints differs from the probability distribution of time intervals between outliers we can use this information to predict time locations of the changes and skip outliers and therefore achieve better TP/FP rates.

BLPA is a new online detection method. It extends the Bayesian Online Changepoint Detector (**BD**) proposed in [4] by embedding into it a Predictive Change Confidence Function

(**PCCF**), which we introduced recently in [5], in order to predict future changepoints in the input data stream, adjust detector's settings dynamically and to reduce FP rate.

In short, BD detector works by recursively estimating posterior probability distribution $P(r_t|\boldsymbol{x}_{1:t}, \theta)$ of the *run length* variable r_t which is a time since the last changepoint. Changepoint is an event when

$$\arg\max_{r_t} P(r_t|\boldsymbol{x}_{1:t}, \theta) = 0$$

The *posterior* distribution is recalculated then every time a new measurement x_t is observed using Bayes' theorem to update parameters of the distributions used to model data and the law of total probability

$$P(r_t|\ \bullet) = \sum_{r_{t-1}} P(r_t|\ r_{t-1},\ \bullet)\ P(r_{t-1}|\ \bullet)$$

to consider all possible run's values in the past.

The *prior* probability of the change $P(r_t = 0|t)$ in BD detector is specified using the constant-value hazard rate h which is an instant prior probability to observe a change and which is supposed to be known before the change detection process starts. The uniform *non-informative* prior does not hold enough information to distinguish outliers and noisy changes from the changepoints. We improve performance of the BD detector by using an *informative* prior distribution in a form of the PCCF function which parameters are the average time interval μ between consecutive changepoints $\langle c_i - c_{i-1}\rangle$ and standard deviation σ. Given current estimate of μ and σ PCCF gives a prior probability $\mathcal{P}(t|\mu, \sigma)$ to observe recurrent changepoint at time t. During the change detection process parameters of the BD detector are adjusted dynamically according the predictions in order to skip possible noisy changes in between changepoints. When a new changepoint is detected (or its location is provided by outer source) parameters (μ, σ) are updated using Bayesian rule and new prediction $\mathcal{P}(t|\mu_{\text{new}}, \sigma_{\text{new}})$ is made.

The paper is organized as follows. In Section II we review related works. In Section III we describe in detail how the Bayesian Change Detector proposed in [4] works. In Section IV we describe **PCCF** function used to predict recurrent changes. In Section V we describe the data model common for the input signal of observations $\langle x_i\rangle$ and for the time intervals between changepoints $\langle c_i - c_{i-1}\rangle$. In the Section VI we describe the **BLPA** algorithm which is a **BD** detector integrated with the **PCCF** function. In the Section VII we describe experimental results demonstrating improved performance of the **BD** detector when integrated with the **PCCF**.

II. Related work

While many change detection methods have been developed [1], [6] for offline and online settings, they typically assume that changes occur at random in time, and are independent from each other. In practice, however, in many industrial applications changes occur with some regularity (e.g. seasonality). Our BLPA approach captures this information from data, and utilizes it for improving the accuracy of a Bayesian online change detection.

In the Bayesian online change detector proposed in [4] and extended in [7] authors model time intervals between change points (run lengths) using the hazard rate. This approach allows to take into account recurrence by tuning single parameter, but it does not allow to distinguish outliers from changes which may appear between them. In [8] data stream volatility, defined as the rate of detected changes, is used to make detector more reactive. We concentrate on the problem of improving change detection by predicting time locations of the changes in the future in order to better distinguish outliers from real changes.

In BD [4] the hazard rate is a constant value assumed to be known in advance. This is not a realistic assumption and this problem has been addressed in [9] where authors proposed an on-line inference procedure to estimate h parameter for the case if hazard rate is unknown and can itself undergo changes while new data arrives. In [10] authors proposed an algorithm which can detect and locate changepoints simultaneously using Bayesian statistics approach. In [11] authors use Gaussian Process model to compute predictive distribution $p(x_{\textbf{new}}|\boldsymbol{x}_{\textbf{old}})$.

Our method is different from these ones because we combine change detection and prediction tasks. We add a second layer (PCCF function) on top of the change detection algorithm allowing to predict future recurrent changes and adjust detectors settings dynamically. This second layer is a change detector itself in the sense that it automatically incorporates changes in underlying distribution of the time intervals between recurrent changes.

In our previous work [5] we demonstrated how to integrate PCCF with the very naive threshold based detector in a heuristic way. The BLPA method we propose here is a more advanced. It integrates PCCF natively into the BD detector using Bayesian statistics framework. BLPA updates both parameters of BD and PCCF sequentially, detects changes, predicts future changes and adjusts parameters of the BD detector according to the predictions in order to skip noisy changes and outliers while detecting changes of interest.

A few other and more remote lines of work relate to our approach via attention to recurrent concept drift [12], [13], [14], predictability of concept drift [15], or change detection with delayed labeling [16]. These approaches are specific to handling concept drift, while our focus is on generic online change detection and its accuracy.

III. Online Bayesian Change Detector (BD)

In this section we describe the Bayesian Online Changepoint Detector proposed in [4]. As we mentioned - to model time occurrences of the changes authors introduce a latent variable run length r_t which is the number of time steps since the most recent change. In Fig. 1 plot **(A)** you can see an illustrating example of the input signal and corresponding run values on plot (B).

On each time step there are two possibilities: either the run length increases $r_t = r_{t-1} + 1$ or changepoint occurs $r_t = 0$.

The conditional prior $P(r_t|r_{t-1})$ of the change is given by a constant-value hazard rate h (Equation 1).

$$p(r_t|r_{t-1}) = \begin{cases} 1-h & \text{if } r_t = r_{t-1}+1 \\ h & \text{if } r_t = 0 \end{cases} \quad (1)$$

The plot **(C)** in Fig. 1 illustrates the message-passing algorithm to compute prior probabilities of the changepoint at any time moment given the boundary condition $P(r_1 = 0) = 1.0$ that change occurred at the moment $t = 1$. Each node (circle) represents a hypothesis about the current run length value. From each node there is a solid line upwards depicting probability of increasing of the run on the next time step (no change) and a dashed line going downwards depicting probability of the change.

At each time step the probability of the changepoint is estimated by calculating posterior probability distribution of the run length value given the data so far observed (Equation 2).

$$P(r_t|\boldsymbol{x}_{1:t}) = \frac{P(r_t, \boldsymbol{x}_{1:t})}{P(\boldsymbol{x}_{1:t})} \quad (2)$$

The joint probability of the run length values and observed so far data can be sequentially computed using recursive procedure in Equation 3 as it is described in [4]:

$$\begin{aligned} P(r_t, x_{1:t}) &= \sum_{r_{t-1}} P(r_t, r_{t-1}, x_{1:t}) = \\ &\sum_{r_{t-1}} P(r_t, x_t \mid r_{t-1}, x_{1:t-1})\, P(r_{t-1}, x_{1:t-1}) = \\ &\sum_{r_{t-1}} P(r_t|r_{t-1}) P(x_t|r_{t-1}, x_t^{(r)}) P(r_{t-1}, x_{1:t-1}) \end{aligned} \quad (3)$$

where $x_t^{(r)} \equiv \langle x_{t-r+1}, \ldots, x_t \rangle$ is input data sub-interval associated with the run length r. Marginal predictive distribution of the new observation x_t is computed using the sum rule:

$$P(x_t|\boldsymbol{x}_{1:t-1}) = \sum_{r_t} P(x_t|r_t, \boldsymbol{x}_t^{(r)}) P(r_t|\boldsymbol{x}_{1:t-1}) \quad (4)$$

IV. PCCF FUNCTION

In this section we show how to compute PCCF function used to predict time locations of the recurrent changes in the future. We consider a discrete case when observations are obtained at the discrete time moments $\langle t \rangle_{t=1}^{T}$ with a constant sampling rate. Probability distribution for the discrete sets is defined using Probability mass function (**Pmf**). As mentioned earlier we assume that changes *re*-occur after 'approximately' equal time intervals. To model time intervals between consecutive changes $\langle c_i - c_{i-1} \rangle$ we use the Gaussian distribution assuming that standard deviation is small enough so that probability to observe the change c_i before c_{i-1} is extremely small.

Definition 1: Changes $\langle c_i \rangle_{i=1}^{k}$ are recurrent if

$$p(c_{i+1} = t \mid \theta^C) = p(c_1 = t - c_i \mid \theta^C), \quad (5)$$

where $\theta^C = (\mu^C, \sigma^C)$, c_1 is the time of the 1^{st} change, c_i is the time of the i^{th} change.

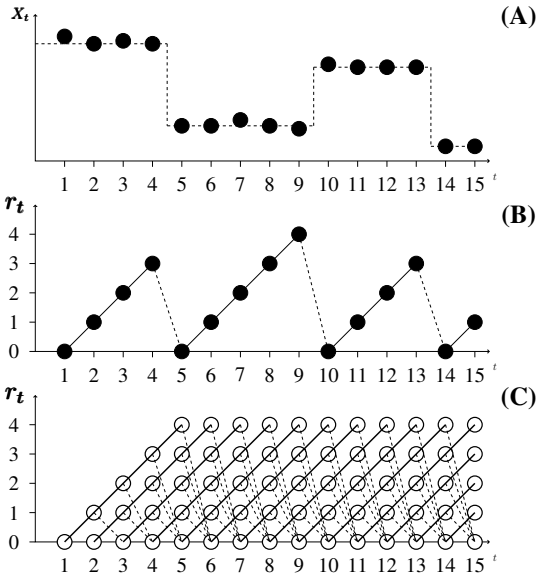

Fig. 1. The changepoint detection problem. (A): Input signal. (B): A particular realization of the run length path corresponding to the actual changepoints locations in the input signal. (C): Directed graph representing all possible run length paths. *The figure is replicated from the illustration in [4].*

This definition corresponds to the generative model defined by Equation 6 in which every next change c_{i+1} happens after time intervals Δ which are samples from the Gaussain distribution $N(\mu^C, \sigma^C)$.

$$c_{i+1} = c_i + \Delta, \quad \text{where } \Delta \sim N(\mu^c, \sigma^c) \quad (6)$$

To predict future changes we introduce the notion of the Predictive Change Confidence Function (**PCCF**) [5].

*Definition 2: **PCCF** is a **Pmf** defined on a discrete set of time moments $\langle t \rangle_{t=1}^{T}$ giving a probability to observe recurrent change $\forall c \in \langle c_i \rangle_{i=1}^{k}$ at the time moment t*

$$\mathcal{P}(c = t|\mu^c, \sigma^c) = \sum_{i=1}^{k} p(c_i = t|\mu^c, \sigma^c) \quad (7)$$

*where $p(c_i = t|\mu^c, \sigma^c)$ is a **Pmf** for an individual change c_i.*

It is important to note that *change*-events $\langle c_i \rangle$ are independent. Every c_i can happen at any moment of time according to its individual **Pmf** $p(c_i = t|\mu^c, \sigma^c)$. Following the sum rule for total probability[1] in order to compute Pmf of c_{i+1} we need to consider all possible time locations of c_i.

$$p(c_{i+1} = t) = \sum_{\tau=i}^{t-1} p(c_{i+1} = t \mid c_i = \tau) p(c_i = \tau). \quad (8)$$

According to the definition 2 PCCF is a sum of individual **Pmf**'s of the changes which might happen till current moment

[1] $P(x) = \sum_y P(x|y)p(y)$

of time

$$\begin{aligned}\mathcal{P}(t) &= \sum_{i=1}^{t}\sum_{\tau=i}^{t-1} p(c_{i+1}=t|c_i=\tau)p(c_i=\tau) \\ &= \sum_{i=1}^{t}\sum_{\tau=i}^{t-1} p(c_1=t-c_i)p(c_i=\tau). \end{aligned} \quad (9)$$

Right side of the Equation 8 is a convolution for the Pmf $p(c_1)$ of the 1^{st} recurrent change and of the Pmf of the change c_i computed in the previous step

$$\begin{aligned} p(c_{i+1}) &= (p(c_1) * p(c_i))[\tau] \\ &= \sum_{\tau=1}^{t-1} p(c_1=t-\tau)p(c_i=\tau). \end{aligned} \quad (10)$$

The convolution of two Gaussian distributions is also Gaussian distribution

$$(p(x|\mu_1,\sigma_1) * p(x|\mu_2,\sigma_2)) = p(x|\mu_1+\mu_2, \sqrt{\sigma_1^2+\sigma_2^2}). \quad (11)$$

PCCF (Eq. 9) can be written as a t-fold convolution

$$\mathcal{P}(t) = \underbrace{(p(c_1) * p(c_1) * \cdots * p(c_1))}_{\text{t}} \quad (12)$$

which is equivalent to the sum

$$\mathcal{P}(t) = \sum_{l=1}^{t} \frac{1}{\sigma\sqrt{2\pi l}} \exp\left(\frac{-(t-l\mu)^2}{2l\sigma^2}\right). \quad (13)$$

The sum 13 describes renewal-reward process [17], [18]. Using the renewal theorem [17] we can calculate the limit of $\mathcal{P}(t)$ when $t \to \infty$

$$L = \lim_{t\to\infty} \sum_{l=1}^{\infty} \frac{1}{\sigma\sqrt{2\pi l}} \exp\left(-\frac{(t-\mu l)^2}{2l\sigma^2}\right) = \frac{1}{\mu}. \quad (14)$$

From Equation 14 follows that PCCF converges to the constant value uniform distribution for large t values. Fig. 2 illustrates two PCCF functions (Equation 13) with parameters $(\mu = 10, \sigma = 2)$ and $(\mu = 15, \sigma = 3)$. Prior and posteriors for

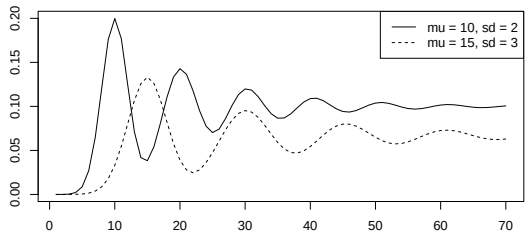

Fig. 2. An example of two Gaussian PCCF functions. The limits are $\frac{1}{\mu^C}$.

the PCCF's parameters are estimated and updated using the procedure described in Section V describing data model.

V. Data model

In this section we describe the data model which we use in BLPA. There are two streams of data to be analysed. The stream of input observations $\langle x_t \rangle$ and the stream of time intervals between changepoints $\langle c_i - c_{i-1} \rangle$. The first stream is used to detect changes and therefore to produce the second stream. The stream of changes maybe updated by the outer sources providing additional information about time location of the changes. E.g. there is a process running in parallel with the main detector which can run additional change-detection processes over collected data to identify locations of the changes in the past more precisely. The stream of changepoints is used to predict future changepoints in order to adjust detector's settings to achieve a better performance. Further, data D is either input data stream of observations $\langle x_t \rangle$ or data stream of time intervals between consecutive changepoints $\langle c_i - c_{i-1} \rangle$. In this section we describe a data model for D common both for input signal and sequence of time intervals between changepoints.

Data D is assumed to be generated by a Gaussian distribution with an unknown mean and variance. We denote elements of D by $\tilde{x}_i \in D$ with mean and variance $(\tilde{\mu}, \tilde{\sigma})$.

A. Prior distributions.

Following the notations in [19], we use a *normal-gamma* prior for $\tilde{\mu}$ and $\tilde{\sigma}$:

$$\tilde{x}_i \sim N(\tilde{\mu}, \tilde{\tau}), \quad \tilde{\tau} = (1/\tilde{\sigma})^2 \quad (15)$$

$$\tilde{\mu} \sim N(\tilde{\mu}_0, \tilde{\kappa}_0\tilde{\tau}) \quad (16)$$

$$\tilde{\tau} \sim Gamma(\tilde{\alpha}_0, \tilde{\beta}_0) \quad (17)$$

where $(\tilde{\alpha}_0, \tilde{\beta}_0, \tilde{\mu}_0, \tilde{\kappa}_0)$ are hyperparameters. The value $\tilde{\tau}$ is also named *precision* [2]. The likelihood of data $D = \langle \tilde{x}_i \rangle$ is

$$P(D|\tilde{\mu}, \tilde{\tau}) = \Big(\frac{\tilde{\tau}}{2\pi}\Big)^{n/2} \exp\Big(-\frac{\tilde{\tau}}{2}\sum_{i=1}^{n}(\tilde{x}_i - \mu)^2\Big) \quad (18)$$

The joint conjugate prior for parameters $(\tilde{\mu}, \tilde{\tau})$ is the defined *normal-gamma* (*NG*) distribution:

$$P(\tilde{\mu}, \tilde{\tau}|\tilde{\mu}_0, \tilde{\kappa}_0, \tilde{\alpha}_0, \tilde{\beta}_0) = N(\tilde{\mu}_0, \tilde{\kappa}_0\tilde{\tau})Gamma(\tilde{\alpha}_0, \tilde{\beta}_0) \quad (19)$$

$$= \frac{1}{Z}\tilde{\tau}^{1/2}\exp\Big(-\frac{\tilde{\kappa}_0\tilde{\tau}}{2}(\tilde{\mu}-\tilde{\mu}_0)^2\Big)\tilde{\tau}^{\tilde{\alpha}_0-1}e^{-\tilde{\tau}\tilde{\beta}_0} \quad (20)$$

$$= \frac{1}{Z}\tilde{\tau}^{\tilde{\alpha}_0-1/2}\exp\Big(-\frac{\tilde{\tau}}{2}[\tilde{\kappa}_0(\tilde{\mu}-\tilde{\mu}_0)^2 + 2\tilde{\beta}_0]\Big) \quad (21)$$

where $Z = \frac{\Gamma(\tilde{\alpha}_0)}{\tilde{\beta}_0^{\tilde{\alpha}_0}}\Big(\frac{2\pi}{\tilde{\kappa}_0}\Big)^{1/2}$ is the normalized factor.

B. Posterior distributions

The posterior can be derived as

$$P(\tilde{\mu}, \tilde{\tau}|D) \propto P(\tilde{\mu}, \tilde{\tau}|\tilde{\mu}_0, \tilde{\kappa}_0, \tilde{\alpha}_0, \tilde{\beta}_0)P(D|\tilde{\mu}, \tilde{\tau}) \quad (22)$$

$$\propto N(\tilde{\mu}_n, \tilde{\kappa}_n\tilde{\tau})Gamma(\tilde{\alpha}_0 + n/2, \tilde{\beta}_n) \quad (23)$$

which is also a *normal-gamma* distribution:

$$P(\tilde{\mu}, \tilde{\tau}|D) = NG(\tilde{\mu}, \tilde{\tau}|\tilde{\mu}_n, \tilde{\kappa}_n, \tilde{\alpha}_n, \tilde{\beta}_n) \quad (24)$$

[2]Further we use σ and τ parameters interchangeably.

with the parameters

$$\tilde{\mu}_n = \frac{\tilde{\kappa}_0}{\tilde{\kappa}_0 + n}\tilde{\mu}_0 + \frac{n}{\tilde{\kappa}_0 + n}\bar{x} \tag{25}$$

$$\tilde{\kappa}_n = \tilde{\kappa}_0 + n \tag{26}$$

$$\tilde{\alpha}_n = \tilde{\alpha}_0 + n/2 \tag{27}$$

$$\tilde{\beta}_n = \tilde{\beta}_0 + \frac{1}{2}\sum_{i=1}^{n}(\tilde{x}_i - \bar{x})^2 + \frac{\tilde{\kappa}_0 n(\bar{x} - \tilde{\mu}_0)^2}{2(\tilde{\kappa}_0 + n)} \tag{28}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i$ is the mean of sampled data. The posterior distribution for $\tilde{\tau}$ is obtained by integrating Equation 24 over μ (See [19]) –

$$p(\tilde{\tau}|D, \tilde{\mu}_0, \tilde{\kappa}_0, \tilde{\alpha}, \tilde{\beta}) \propto$$
$$Gamma(\tilde{\alpha}+n/2, \tilde{\beta}+\frac{1}{2}\sum_{i=1}^{n}(\tilde{x}_i-\bar{x})^2+\frac{\tilde{\kappa}\tilde{\kappa}_0}{2(\tilde{\kappa}+\tilde{\kappa}_0)}(\bar{x}-\tilde{\mu}_0)^2)) \tag{29}$$

Given the updated parameters $\theta = (\alpha_0, \beta_0, \mu_0, \kappa_0)$ using the rules 28 , the predictive distribution for a new data x_{new} is

$$p(x_{\text{new}}|\boldsymbol{x}, \mu, \kappa, \alpha, \beta) = \int p(x_{\text{new}}|\mu, \tau)p(\tau|\boldsymbol{x}, \mu_0, \kappa_0, \alpha, \beta)d\tau \tag{30}$$

where

$$p(x_{\text{new}}|\mu, \tau) = (\frac{\tau}{2\pi})^{1/2}e^{-\frac{\tau}{2}(x-\mu)^2}d\tau \tag{31}$$

and $p(\tau|\boldsymbol{x}, \mu_0, \kappa_0, \alpha, \beta)$ is given by 29. Integral 30 is a Pearson type VII distribution (Equation 32) which is equivalent of the non-standardized Student's t-distribution.

$$p(x_{\text{new}}) = \frac{1}{\alpha B(m - 1/2, 1/2)}\Big(1 + \Big(\frac{x_{n+1} - \lambda}{\alpha}\Big)^2\Big)^{-m} \tag{32}$$

where

$$m = \alpha_0 + (n+1)/2 \tag{33}$$

$$\alpha = A\sqrt{\sum_{i=1}^{n}x_i^2 + \kappa_0\mu_0^2 - \frac{(\sum_{i=1}^{n}x_i + \mu_0\kappa_0)^2}{n + \kappa_0} + 2\beta_0} \tag{34}$$

$$A = \sqrt{\frac{n + 1 + \kappa_0}{n + \kappa_0}} \tag{35}$$

$$\lambda = \frac{\sum_{i=1}^{n}x_i + \mu_0\kappa_0}{n + \kappa_0} \tag{36}$$

Predictive distribution 4 in case of this data model is given by Equation 32. Please see detailed calculations in the Appendix.

VI. BLPA change detector

The BLPA method is a combination of BD detector and PCCF predictive function. Particularly when we compute the joint probability $P(r_t, \langle x_j\rangle_{j=1}^{t})$ and after that when computing the run-length distribution $P(r_t|\langle x_j\rangle_{j=1}^{t})$ we multiply these probabilities by the prior probability of the changes given by PCCF for the moment t. The BLPA method is depicted in Algorithm 1, in which:

- Lines 1-4: Set initial parameters values for the probability distribution of the data D.
- Line 5: Compute PCCF using initial values of the parameters (Equation 9).
- Line 7: Collect a new measurement.
- Line 8: Compute predictive distribution using Equation 32.
- Line 9-10: Compute change probabilities and 'growth' probabilities of the run length.
- Line 11: Compute posterior probabilities of run lengths (changes).
- Line 12: Update parameters of the probability distributions for the data D using Equations 28.
- Lines 13-16: Find the most likely position of the last changepoint, update PCCF parameters and recalculate PCCF.

Algorithm 1 LPA-detector pseudocode

1: $\boldsymbol{\theta} \leftarrow (\mu_0, \kappa_0, \alpha_0, \beta_0)$
2: $\boldsymbol{\theta^C} \leftarrow (\mu_0^C, \kappa_0^C, \alpha_0^C, \beta_0^C)$
3: $\boldsymbol{\theta = \theta_0}$ ▷ Init sig. params
4: $\boldsymbol{\theta^C = \theta_0^C}$ ▷ Init PCCF params
5: $\boldsymbol{\langle H_j\rangle_{j=1}^{T}} = Pccf(\boldsymbol{\theta_0^C})$ ▷ Predict changes (Initial)
6: **for** t=1:T **do**
7: $\boldsymbol{x} \leftarrow [\boldsymbol{x}, x_t]$ ▷ Observe new datum
8: $\boldsymbol{\pi_t} = P(x_t|\boldsymbol{\theta})$ ▷ Predicitve distribution
9: $P(r_t = r_{t-1} + 1, \boldsymbol{x}) = P(r_{t-1}, x_{1:t-1})\boldsymbol{\pi_t}(1 - \boldsymbol{H_{t-1}})$
10: $P(r_t = 0, \boldsymbol{x}) = \boldsymbol{H_{t-1}}\sum_{r_{t-1}}P(r_{t-1}, x_{1:t-1})\boldsymbol{\pi_t}$
11: $P(r_t|\boldsymbol{x}) = P(r_t, \boldsymbol{x})/P(\boldsymbol{x})$ ▷ Run length Distrib
12: $\boldsymbol{\theta} \leftarrow$ Update($\boldsymbol{\theta}$) ▷ Update parameters
13: **if** $(\arg\max_{r_t} p(r_t|\boldsymbol{x}, \boldsymbol{\theta}) = 0)$ **then**
14: $\boldsymbol{\theta^C} \leftarrow$ Update($\boldsymbol{\theta^C}$)
15: $\boldsymbol{\langle H_j\rangle_{j=t}^{T}} = Pccf(\boldsymbol{\theta^C})$
16: **end if**
17: **end for**

VII. Experiments

We performed experiments with artificially generated and real data sets. To measure the performance of the change detector we can consider it as as a binary classifier assigning labels 'change'/'not change' to the incoming observations x_t. If $\mathbf{e}_t^+$ is the 'change' label assigned at the moment t and $\mathbf{e}_t^-$ is the label 'not change' assigned at t then Then True Positive (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) events can be defined as follows:

- $\mathbf{e}_t^+$ is TP if $\exists c_i : t - c_i < \delta$, and FP if $\nexists c_i : t - c_i < \delta$
- $\mathbf{e}_t^-$ is FN if $\exists c_i : t - c_i < \delta$, and TN if $\nexists c_i : t - c_i < \delta$

The *performance of the change detector* is defined by TP/FP rates and by the average delay δ of the detection.

A. Artificial data

In the simulation we generated 200 signals with 10 recurrent changes in the mean value for each hazard-rate value h varied in the interval from 50 to 300 by the step 15. Average distance between changes was set to $\mu = 100$ with the standard deviation $\sigma = 10$. Results are depicted in Fig. 3. FP rate

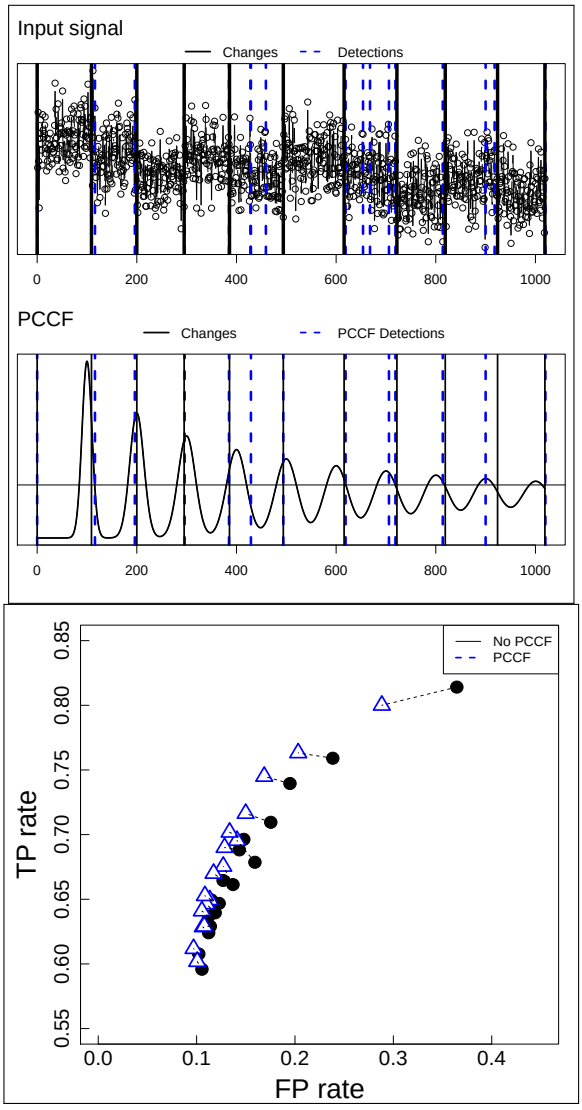

Fig. 3. Experimental results for simulated data streams with recurrent changes. On the top plot - an example of the generated input signal. Vertical solid lines depict changepoints to be detected. Dashed lines on the plot with the signal depict moments when detector **without** PCCF alarmed changes. Bottom plot - ROC curves. Blue triangles depict performance of the detector equipped with PCCF. Black dots - performance without PCCF. FP rate is reduced while keeping the same TP rate.

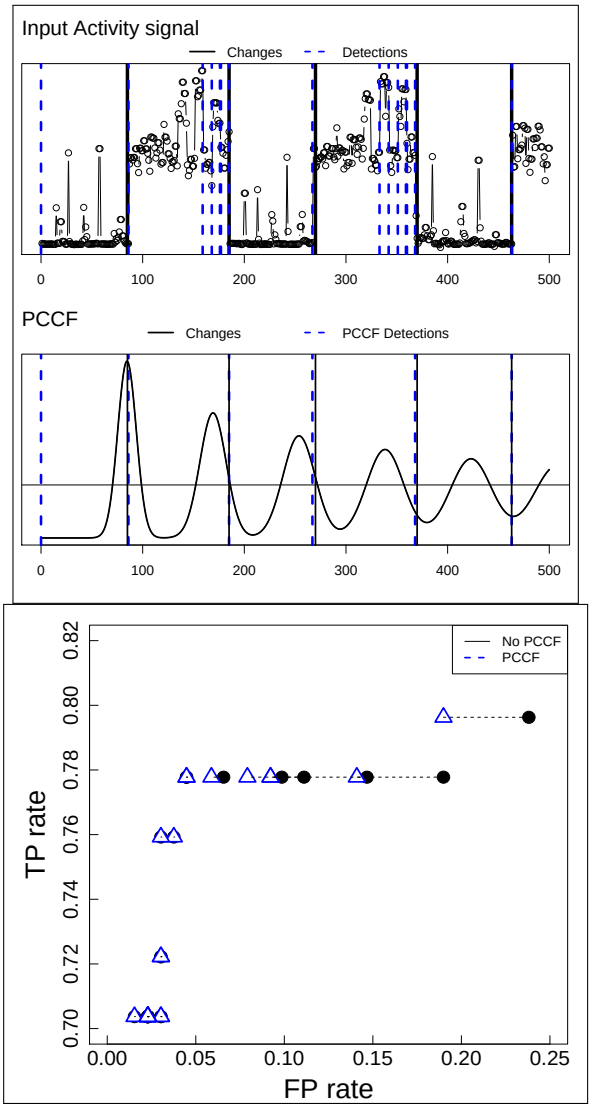

Fig. 4. Experimental results for the 'Activity recognition' signal. On the top plot - illustrating example of the signal and corresponding PCCF function. Vertical solid lines depict changepoints to be detected. Dashed lines on the plot with the signal depict moments when detector **without** PCCF alarmed changes. Dashed lines on the plot **with** PCCF show time moments when the detector with PCCF alarmed changes. Bottom plot - ROC curves. Blue triangles - performance of the BD with PCCF.

is decreased while not reducing TP rate. In the worst cases the performance of both detectors is similar.

B. Human Activity (HA) signal

In the second experiment we used the Human activity data set [20] which contains sensor measurements from people performed 6 types of activities: three static postures (standing, sitting, lying) and three dynamic activities (walking, walking downstairs and walking upstairs). We detected changes in the signal caused by transitions from one set activities to another. Results are depicted in Fig. 4. FP rate is decreased when BD detector is used with the PCCF function.

VIII. Conclusion

We proposed the method to improve performance of the Bayesian Online Changepoint detector (BD) for the data streams with recurrent changes by embedding into it the Predictive Confidence Change Function (PCCF). While observing a new data both BD detector's and PCCF's parameters are adjusted in a uniform way to the changing conditions using the same Bayesian update procedures constituting a two-layer adaptive change detection/prediction method BLPA. In the experiments with real and artificial data sets we demonstrated that Bayesian detector equipped with PCCF performs better in terms of TP/FP rates than the detector without PCCF.

REFERENCES

[1] I. V. Nikiforov and M. Basseville, "Detection of Abrupt Changes," 1993.

[2] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," *Mach. Learn.*, vol. 23, no. 1, pp. 69–101, 1996.

[3] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, pp. 44:1–44:37, 2014.

[4] D. J. M. Ryan Prescott Adams, "Bayesian online changepoint detection," 2007.

[5] A. Maslov, M. Pechenizkiy, I. Žliobaitė, and T. Kärkkäinen, "Modelling recurrent events for improving online change detection," in *SIAM International Conference on Data Mining (SDM16)*, 2016.

[6] A. S. Polunchenko and A. G. Tartakovsky, "State-of-the-Art in Sequential Change-Point Detection," *Methodology and Computing in Applied Probability*, vol. 14, no. 3, pp. 649–684, Oct. 2011.

[7] R. C. Wilson, M. R. Nassar, and J. I. Gold, "Bayesian online learning of the hazard rate in change-point problems," *Neural computation*, vol. 22, no. 9, pp. 2452–2476, 2010.

[8] D. Huang, Y. S. Koh, G. Dobbie, and R. Pears, "Detecting volatility shift in data streams," in *ICDM'2014*, pp. 863–868.

[9] R. C. Wilson, M. R. Nassar, and J. I. Gold, "Bayesian online learning of the hazard rate in change-point problems," *Neural Comput.*, vol. 22, no. 9, pp. 2452–2476, Sep. 2010. [Online]. Available: http://dx.doi.org/10.1162/NECO_a_00007

[10] A. B. Downey, "A novel changepoint detection algorithm," 2008. [Online]. Available: http://arxiv.org/abs/0812.1237v1

[11] Y. Saatçi, R. D. Turner, and C. E. Rasmussen, "Gaussian process change point models," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 927–934.

[12] J. Gama and P. Kosina, "Learning about the learning process," in *Proc. of IDA'11*, pp. 162–172.

[13] J. B. Gomes, M. M. Gaber, P. A. C. Sousa, and E. M. Ruiz, "Mining recurring concepts in a dynamic feature space," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 25, no. 1, pp. 95–110, 2014.

[14] J. B. Gomes, P. A. C. Sousa, and E. M. Ruiz, "Tracking recurrent concepts using context," *Intell. Data Anal.*, vol. 16, no. 5, pp. 803–825.

[15] H. Ang, V. Gopalkrishnan, I. Zliobaite, M. Pechenizkiy, and S. Hoi, "Predictive handling of asynchronous concept drifts in distributed environments," *IEEE Trans. on Knowl. and Data Eng.*, vol. 25, pp. 2343–2355, 2013.

[16] I. Zliobaite, "Change with delayed labeling: When is it detectable?" in *ICDM 2010 Workshops*, pp. 843–850.

[17] D. R. Cox, *Renewal theory*. Methuen, 1962, vol. 58.

[18] W. Feller, *An introduction to probability theory and its applications: volume I*. John Wiley & Sons London-New York-Sydney-Toronto, 1968, vol. 3.

[19] M. I. Jordan, "Chapter 9. the exponential family: Conjugate priors." [Online]. Available: https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/other-readings/

[20] J.-L. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, and D. Anguita, "Transition-aware human activity recognition using smartphones," *Neurocomputing*, vol. 171, pp. 754–767, 2016.

APPENDIX

Assuming Gaussian distribution

$$p(x|\mu,\tau)=\frac{\sqrt{\tau}}{\sqrt{2\pi}}e^{-\frac{\tau}{2}(x-\mu)^2}$$

Probability $p(x_{n+1}|x_1,...,x_n)$ can be found as

$$p(x_{n+1}|x_1,...,x_n) \tag{37}$$
$$=\frac{\int_{\tau\in\mathbb{R}^+}\int_{\mu\in\mathbb{R}}p(x_1,....,x_{n+1}|\mu,\tau)p(\mu,\tau)d\mu d\tau}{\int_{\tau\in\mathbb{R}^+}\int_{\mu\in\mathbb{R}}p(x_1,....,x_n|\mu,\tau)p(\mu,\tau)d\mu d\tau}.$$

The value $p(x_1,....,x_n|\mu,\tau)p(\mu,\tau)$ is

$$\left(\frac{\sqrt{\tau}}{\sqrt{2\pi}}\right)^n\times\exp\left(-\frac{\tau}{2}\sum_{i=1}^n(x_i-\mu)^2\right)\tau^{\alpha_0-1/2} \tag{38}$$
$$\times\exp\left(-\frac{\tau}{2}\left(\kappa_0(\mu-\mu_0)^2+2\beta_0\right)\right).$$

An expression in the exponent is equivalent to

$$-\frac{\tau(n+\kappa_0)}{2}\sum_{i=1}^n\left(\mu-\frac{\sum_{i=1}^n x_i+\mu_0\kappa_0}{n+\kappa_0}\right)^2 \tag{39}$$
$$-\frac{\tau}{2}\left(\sum_{i=1}^n x_i^2+\kappa_0\mu_0^2-\frac{\left(\sum_{i=1}^n x_i+\mu_0\kappa_0\right)^2}{n+\kappa_0}+2\beta_0\right),$$

from where

$$p(x_1,....,x_n|\mu,\tau)p(\mu,\tau) \tag{40}$$
$$=\left(\frac{1}{\sqrt{2\pi}}\right)^{n-1}\frac{\Gamma(\alpha_0+n/2)}{\sqrt{n+\kappa_0}}\left(\frac{\hat{b}_n}{2}\right)^{-\alpha_0+n/2}$$
$$\times p_{Gamma(\alpha_0+n/2,\hat{a}_n/2)}(\tau)\ p_{\mathcal{N}(\hat{\mu}_n,\hat{\sigma}_n^2)}(\mu)$$

where

$$\hat{a}_n=\left(\sum_{i=1}^n x_i^2+\kappa_0\mu_0^2-\frac{\left(\sum_{i=1}^n x_i+\mu_0\kappa_0\right)^2}{n+\kappa_0}+2\beta_0\right), \tag{41}$$

$$\hat{\mu}_n=\frac{\sum_{i=1}^n x_i+\mu_0\kappa_0}{n+\kappa_0},\ \hat{\sigma}_n^2=\frac{1}{\tau(n+\kappa_0)}.$$

Therefore

$$\int_{\tau\in\mathbb{R}^+}\int_{\mu\in\mathbb{R}}p(x_1,....,x_{n+1}|\mu,\tau)p(\mu,\tau)d\mu d\tau \tag{42}$$
$$=\left(\frac{1}{\sqrt{2\pi}}\right)^{n-1}\frac{\Gamma(\alpha_0)}{\sqrt{n+\kappa_0}}\left(\frac{\hat{a}_n}{2}\right)^{-(\alpha_0+n/2)}$$

and integral 37 can be expressed as

$$\frac{\sqrt{n+\kappa_0}}{\sqrt{n+1+\kappa_0}\ B(\alpha_0+n/2,1/2)}\frac{\hat{a}_n^{\alpha_0+n/2}}{\hat{a}_{n+1}^{\alpha_0+(n+1)/2}} \tag{43}$$
$$=\frac{\sqrt{n+\kappa_0}}{\sqrt{n+1+\kappa_0}\ B(\alpha_0+n/2,1/2)}\left(\frac{\hat{a}_{n+1}}{\hat{a}_n}\right)^{-(\alpha_0+(n+1)/2)}\hat{a}_n^{-1/2}.$$

Noting that

$$\frac{\hat{a}_{n+1}}{\hat{a}_n}=1+\frac{n+\kappa_0}{\hat{a}_n(n+1+\kappa_0)} \tag{44}$$
$$\times\left(x_{n+1}^2-\frac{2x_{n+1}\left(\sum_{i=1}^n x_i+\mu_0\kappa_0\right)}{n+\kappa_0}+\frac{\left(\sum_{i=1}^n x_i+\mu_0\kappa_0\right)^2}{(n+\kappa_0)^2}\right)$$

from where

$$p(x_{n+1}|x_1,...,x_n) \tag{45}$$
$$=\frac{1}{\hat{b}_nB(\alpha_0+n/2,1/2)}\left(1+\frac{(x_{n+1}-\lambda)^2}{\hat{b}_n^2}\right)^{-(\alpha_0+(n+1)/2)},$$

where coefficients are

$$\hat{b}_n=\frac{\sqrt{(n+1+\kappa_0)\hat{a}_n}}{\sqrt{(n+\kappa_0)}},\ \lambda=\frac{\sum_{i=1}^n x_i+\mu_0\kappa_0}{n+\kappa_0} \tag{46}$$